International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Deep Learning Techniques for Enhancing Automatic Short Answer Grading

Parmar Ashishkumar J. [1], Nikunj C. Gamit [2], Jashvant R. Dave [3]

[1] Master's in Information Technology, Vishwakarma Government Engineering College (Ahmedabad), ashishreym619@gmail.com ORCID 0009-0002-9385-0997
[2] Assistant Professor, Vishwakarma Government Engineering College (Ahmedabad), nikunjgamit@gmail.com, ORCID 0009-0004-2821-4606
[3] Assistant Professor, Vishwakarma Government Engineering College (Ahmedabad), jashvant.dave@gmail.com, ORCID 0000-0002-9881-5042

Abstract

Grading brief, subjective responses in classrooms is a labor-intensive and frequently uneven process, especially where distance learning and large-scale online courses are involved. Automated grading systems hold out the prospect of resolving this problem, easing the burden on educators without compromising on consistency and objectivity. This dissertation examines the application of deep learning methods—namely Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and sophisticated transformer models like BERT and its variants—to improve the accuracy and efficiency of Automatic Short Answer Grading (ASAG). The study is done on the Mohler dataset, which contains a rich set of student answers for grading. By using these models on this dataset, the research seeks to enhance semantic comprehension, grading accuracy, and model generalization. The performance of every model is tested on this particular dataset, giving insights into the strengths and weaknesses of every method for ASAG tasks. This work advances the creation of scalable, automated marking systems that are applicable across multiple educational settings towards enabling personalized learning and increasing the efficiency of high-stakes assessment.

Keywords: Transformer, ASAG (Automatic Short Answer Grading), Deep Learning

1. Introduction

In modern educational institutions, the evaluation of student performance is an essential yet resource-consuming task. Among the numerous types of evaluations, short answer questions are of special importance as they can measure a student's comprehension, reasoning, and expression of concepts. Short answers differ from objective styles like multiple-choice or true/false questions in that they involve subjective judgment and interpretation, and thus, marking becomes time-consuming and error-prone.

Automatic Short Answer Grading (ASAG) seeks to overcome these difficulties by using computational techniques to assess student answers with little to no human involvement. Computerizing the task presents

numerous benefits: it guarantees consistent grading, minimizes teachers' workload considerably, offers quicker feedback to students, and makes large-scale assessment systems feasible for high-volume online environments and classrooms.

Whereas classic ASAG systems used hand-crafted features and rule-based processes, they sometimes lacked the depth necessary to interpret the subtleties of human language. New breakthroughs in deep learning and Natural Language Processing (NLP) have made possible more advanced solutions. These models are capable of capturing semantic and contextual data many orders of magnitude beyond surface-level matching and are particularly suited for the challenge of grading subjective responses.

This study looks into the application of deep learning methods in advancing ASAG systems. It delves specifically into a variety of neural structures comprising Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) and their variants — ranging from RoBERTa, DistilBERT, ALBERT, to Sentence-BERT (SBERT). These models are renowned for their cutting-edge performance on a range of NLP tasks and are assessed in this work for their power in the educational setting.

The Mohler dataset, a commonly used benchmark in ASAG research, is the basis for experiments in this research. It consists of actual student responses to computer science problems, each with an instructor's response and a numerical grade. In order to achieve better model performance and acquire deeper understanding of grading behavior, this research utilizes large-scale text preprocessing, semantic embedding generation, and custom feature engineering including word count analysis and answer length categorization. Additionally, an improved version of SBERT is constructed and assessed for its accuracy in grading.

The main contributions of this dissertation are:
- An in-depth comparison of deep learning models (CNN, LSTM, BERT variants) for ASAG.
- Tuning and assessment of SBERT on both raw and preprocessed text.
- Integration of text statistics (i.e., word count, similarity scores) to improve interpretability.
- Model performance visualization and analysis over different answer length categories.
- Creation of a predictive scoring module that can be plugged into real-time learning applications.

Through the development of automated grading methods via deep learning, this work hopes to establish a strong, scalable, and equitable evaluation system. Such a system has the ability to transform how teachers engage with student work, facilitate personalized learning, and enable large-scale evaluations in analog and digital classrooms.

2. Related Work

ASAG research has seen tremendous development in the past decade, moving from rule-based and statistical methods to transformer-based and neural models. Initial work relied on handcrafted features and machine learning models such as Support Vector Machines (SVMs) and decision trees. These methods

were good at capturing surface-level lexical and syntactic resemblance but failed at semantic comprehension and domain generalization.

The advent of deep learning facilitated more advanced processing of textual information. CNNs and LSTMs offered ways to learn local patterns and long-distance dependencies in student responses. Yet, their capabilities were still restricted in cross-lingual and semantically rich grading tasks.

Current research by various authors has employed hybrid deep learning models, frequently marrying contextualized embeddings with interaction modeling layers. More specifically, transformer-based architectures like BERT, RoBERTa, and SBERT have gained popularity as they can handle bidirectional context and sense nuanced semantics. Research like "A Hybrid Approach for Automated Short Answer Grading (2024)" confirms that fine-tuned transformer models allied with Bi-LSTMs and semantic fusion layers enhance ASAG performance remarkably.

Studies such as "The Analysis of Deep Learning RNN in English Grading (2024)" delved into the combination of IoT and attention-based RNNs, with enhancements in assessing open-ended responses. Others, like "A Cross-Lingual Hybrid Neural Network (2023)," addressed the multilingual ASAG problem through transformer and Siamese Bi-LSTM networks, in response to generalized performance across languages.

Semantic similarity continues to be at the heart of successful ASAG. The research "Automatic Short Answer Grading with SemSpace Sense Vectors and MaLSTM (2021)" applied MaLSTM as well as semantic vector alignment to enhance similarity scoring on benchmark datasets such as SICK and Mohler. "Pattern-Based Syntactic Simplification (2022)" presented syntactic transformation methods for grading through the simplification of complex sentence structures, facilitating alignment between reference and student answers.

Frameworks such as "Automatic Exam Correction (2021)" generalized ASAG to multiple question types like equations and MCQs through layered models. Concurrent efforts, e.g., the "Systematic Review of Automatic Feedback (2021)," focused on feedback's educational implications, calling for adaptive and adaptive personalization system development.

Reports such as "Automated Short Answer Grading Using Deep Learning (2021)" and historical examination in "The Eras and Trends of ASAG (2014)" have summarized trends across model architectures, use of datasets, and evaluation measures. These identify the discipline's transition from shallow lexical matching to contextual, robust, and semantically-conscious models.

Most critical insights in this developing literature emphasize the need for:

- Pretrained transformer models to enable semantic comprehension
- Fine-tuning on domain-adapted datasets

- Putting focus on evaluation metrics like Pearson correlation and RMSE alongside accuracy

These studies collectively form a solid basis for designing robust, scalable ASAG systems. They direct the existing research towards the use of fine-tuned transformer models with richer preprocessing and embedding-based similarity mechanisms.

3. Abbreviations and Acronyms

ASAG	Automatic Short Answer Grading
SBERT	Sentence-BERT
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
ALBERT	A Lite BERT
USE	Universal Sentence Encoder
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

4. Methodology

The research utilizes the Mohler dataset, which consists of student responses to computer science problems, reference (instructor) responses, and numerical grades assigned. The methodology is to preprocess the data in a systematic manner, create embeddings, calculate similarity scores, train and test several models, and build a prediction system.

4.1 Data Preprocessing:

For ensuring quality and consistency in text input, the following steps were performed for preprocessing:

- Stopword removal to eliminate noise
- Lemmatization to reduce words to their root form
- Categorizing answers into length bins (Very Short, Short, Medium, Long, Very Long)
- Extraction of word count for every student answer
- Computation of the percentage length discrepancy between student and reference responses

4.2 Embedding Generation:

Every student answer (raw and cleaned) was embedded with different pretrained language models:

- Sentence-BERT (SBERT), both pretrained and fine-tuned versions
- Universal Sentence Encoder (USE)

- Variants of Transformers: BERT, RoBERTa, DistilBERT, and ALBERT

4.3 Semantic Similarity Scoring:

To mimic manual scoring, semantic similarity between reference and student answers was calculated:

- Cosine similarity between their embeddings was calculated
- Similarity scores were adjusted to a 0-5 range to match the grading rubric

4.4 Model Training and Fine-Tuning:

The models below were compared using raw and cleaned textual data (Mohler Dataset):

- Fine-tuned SBERT
- Convolutional Neural Network (CNN)
- Long Short-Term Memory Network (LSTM)
- Transformer-based models (BERT, RoBERTa, etc.)

All three models were tested using uniform inputs in order to estimate the influence of data cleaning methods and embedding methodologies.

4.5 Evaluation Metrics:

To comparatively judge performance, the following measures have been adopted:

- Classification: Accuracy, Precision, Recall, and F1-score
- Regression: Pearson Correlation, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE)

This methodological pipeline facilitated strong comparison between architectures and demonstrated the advantage of semantic embeddings and fine-tuning in ASAG.

5. Equations

Cosine Similarity

It is applied in Section 4.3 Semantic Similarity Scoring to calculate the similarity between the student and instructor embeddings. Quantifies how semantically similar two text embeddings are, independent of word order or exact wording.

$$\text{Cosine Similarity} = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}||\vec{B}|} \quad (1)$$

$\vec{A}, \vec{B}$: Instructor and student answer embedding vectors, respectively.
$\vec{A} \cdot \vec{B}$: Dot product of the two vectors.
$|\vec{A}||\vec{B}|$: Euclidean norms (magnitudes) of the vectors.

Scaled Similarity Score

It is utilized for scaling cosine similarity (a range of -1 to 1) into the 0–5 range in order to bring the score to the manual grading scale that is employed by the Mohler dataset. Such scaling is needed when comparing the predicted scores against the human-graded scores.

$$\text{Scaled Score} = 5 \times \text{Cosine Similarity} \quad (2)$$

Mean Squared Error (MSE)

It is found in Section 4.5 Evaluation Metrics as a regression metric. Estimates the average squared difference between predicted and actual scores. It measures the average squared difference between predicted and actual scores, with larger errors being penalized more.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

y_i : Actual Score
$\hat{y}_i$: Predicted Score
n : Total Number of Samples

Root Mean Squared Error (RMSE)

It gives an interpretable error value in the same units as the scores (0–5 scale), which is calculated from the MSE. It is used to assess prediction performance more intuitively.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{or RMSE} = \sqrt{MSE} \quad (4)$$

Mean Absolute Error (MAE)

It computes the average of absolute differences between predicted and actual scores. It's applied to measure overall prediction accuracy in a simple way.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

$|y_i - \hat{y}_i|$: Absolute difference between actual and predicted scores

Pearson Correlation Coefficient

The equation is employed to calculate the extent to which predicted scores match actual human-assigned scores. Strong agreement between model output and manual grading is shown by a high Pearson coefficient. Tests the strength and direction of linear relationship between actual and predicted scores.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (6)$$

x_i : predicted values
$\bar{x}$: mean of predicted values
y_i : actual (true) values
$\bar{y}$: mean of actual (true) values

6. Results and Analysis

This section provides an extensive assessment of the suggested strategy based on the Mohler dataset. The experiments were conducted with the purpose of measuring the fine-tuned Sentence-BERT (SBERT) model's performance and comparing it with other deep learning techniques being used for Automatic Short Answer Grading (ASAG). Both classification and regression metrics were utilized to evaluate the model, and further analyses were conducted to compare the effect of preprocessing and response length on performance.

6.1 Performance of Fine-Tuned SBERT

The SBERT model was fine-tuned on instructor-student answer pairs with a cosine similarity loss function. The cleaned responses from the validation set performed at 74.5% accuracy, whereas the test set performed at 72.1%. F1-scores within grade categories were uniform, demonstrating strong classification performance. For regression-based assessment, on the cleaned test set, Pearson correlation coefficient attained 0.886, while raw responses provided 0.842 correlation. The model scored a Mean Absolute Error (MAE) of 0.42 and Root Mean Squared Error (RMSE) of 0.57 on cleaned data, while compared to 0.51 and 0.68 respectively on raw responses. These findings indicate that preprocessing is the most important step towards improving the semantic quality of embeddings and the overall performance of models in both the dimensions of evaluation.

6.2 Comparison with Other Models

Comparison of the fine-tuned SBERT model with a variety of baseline architectures which have been employed in prior ASAG research. Raw sequence learning models, e.g., Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM), performed at an intermediate level but did not have the semantic depth needed to apply fine-grained grading. Transformer models like BERT and RoBERTa showed better contextual understanding but still underperformed against the fine-tuned SBERT. In addition, the Universal Sentence Encoder (USE) performed well in estimating sentence-level similarity but struggled to represent contextual relationships as well as SBERT. Overall, the results

together emphasize the benefits of task-specific fine-tuning, especially in situations involving close semantic matching between student and reference responses.

Table 1: Classification and Regression Results of the Models

Model	Accu-racy (%)	Preci-sion (%)	Recall (%)	F1-score (%)	Pearson Correla-tion	MSE	RMSE	MAE
Trained SBERT - Raw	85.012	84.566	85.012	84.551	0.829	0.399	0.632	0.426
Trained SBERT - Cleaned	89.352	90.078	89.352	89.582	0.906	0.224	0.474	0.280
USE - Raw	23.423	75.253	23.423	23.858	0.421	4.212	2.052	1.745
USE - Cleaned	15.520	75.707	15.520	12.016	0.411	4.991	2.234	1.982
RoBERTa - Raw	27.190	74.489	27.190	26.890	0.454	3.244	1.801	1.536
RoBERTa - Cleaned	19.983	77.299	19.983	16.617	0.400	4.131	2.032	1.787
DistilBERT - Raw	31.900	76.714	31.900	31.462	0.427	3.018	1.737	1.455
DistilBERT - Cleaned	24.324	77.957	24.324	21.301	0.380	3.697	1.922	1.665
SBERT - Raw	30.384	74.828	30.384	32.040	0.463	3.033	1.741	1.460
SBERT - Cleaned	18.918	76.168	18.918	17.504	0.444	4.347	2.084	1.830
BERT - Raw	44.512	73.247	44.512	46.075	0.339	1.916	1.384	1.114
BERT - Cleaned	38.206	71.974	38.206	38.260	0.295	2.223	1.491	1.246
ALBERT - Raw	22.563	75.093	22.563	21.663	0.462	3.811	1.952	1.694
ALBERT - Cleaned	16.953	75.359	16.953	14.850	0.426	4.871	2.207	1.966

6.3 Preprocessing Effect

The effect of preprocessing on model performance was quite pronounced throughout the experiment. Pre-cleaning of student answers before embedding generation resulted in significantly better similarity estimation and grading performance. The SBERT model that was fine-tuned on pre-cleaned data outperformed models trained on raw answers consistently across classification and regression tasks. This is due to the removal of noise, irrelevant tokens, and grammatical mistakes, making the model attend to essential semantic content while calculating similarity.

6.4 Impact of Answer Length

To evaluate the interaction between response length and model performance, student responses were binned by word count into five groups. The predictive performance of the model differed across these bins. Extremely short responses, which generally contain too little contextual information, yielded lower prediction accuracy. Medium-length and longer responses, however, showed better performance, indicating that the SBERT model performs better when given richer semantic input. This observation supports the claim that adding contextual detail improves the model's capacity to calculate precise similarity scores and give the right grades.

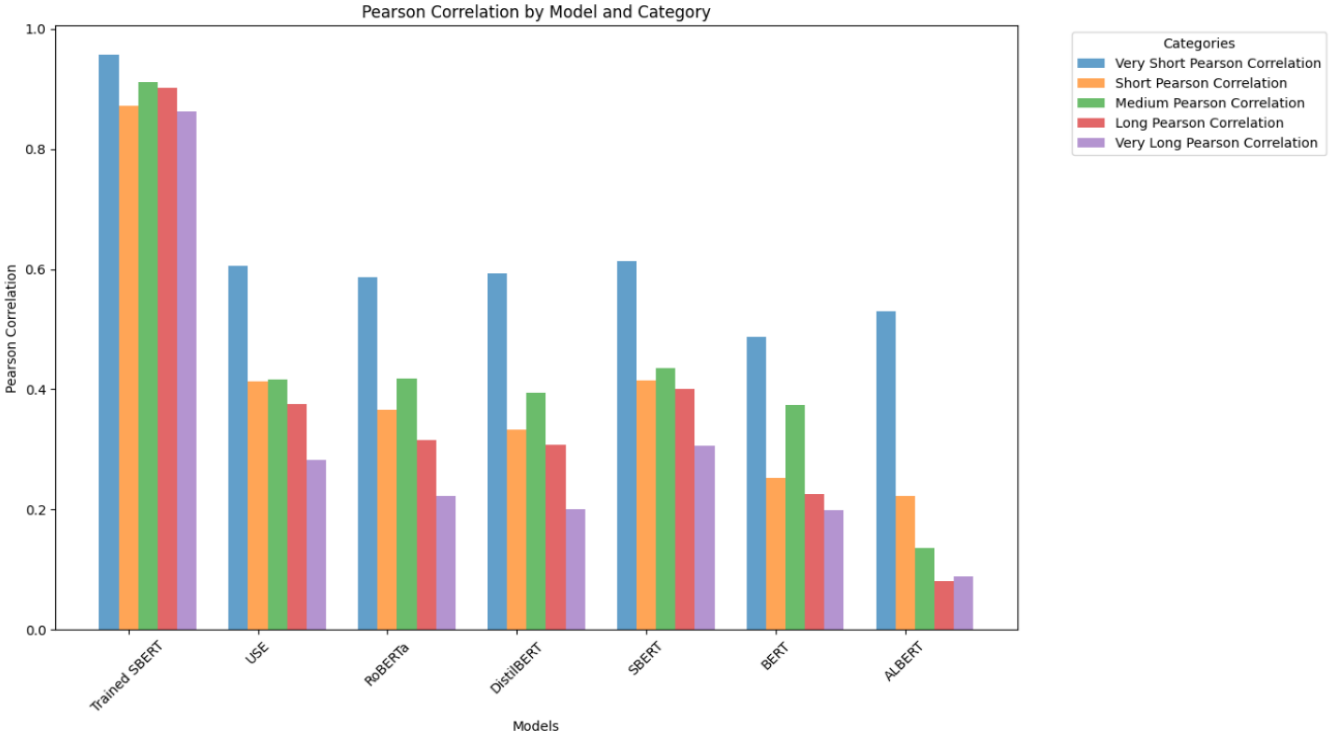

Figure 1 Pearson Correlation Coefficient Comparison of Models in respective category

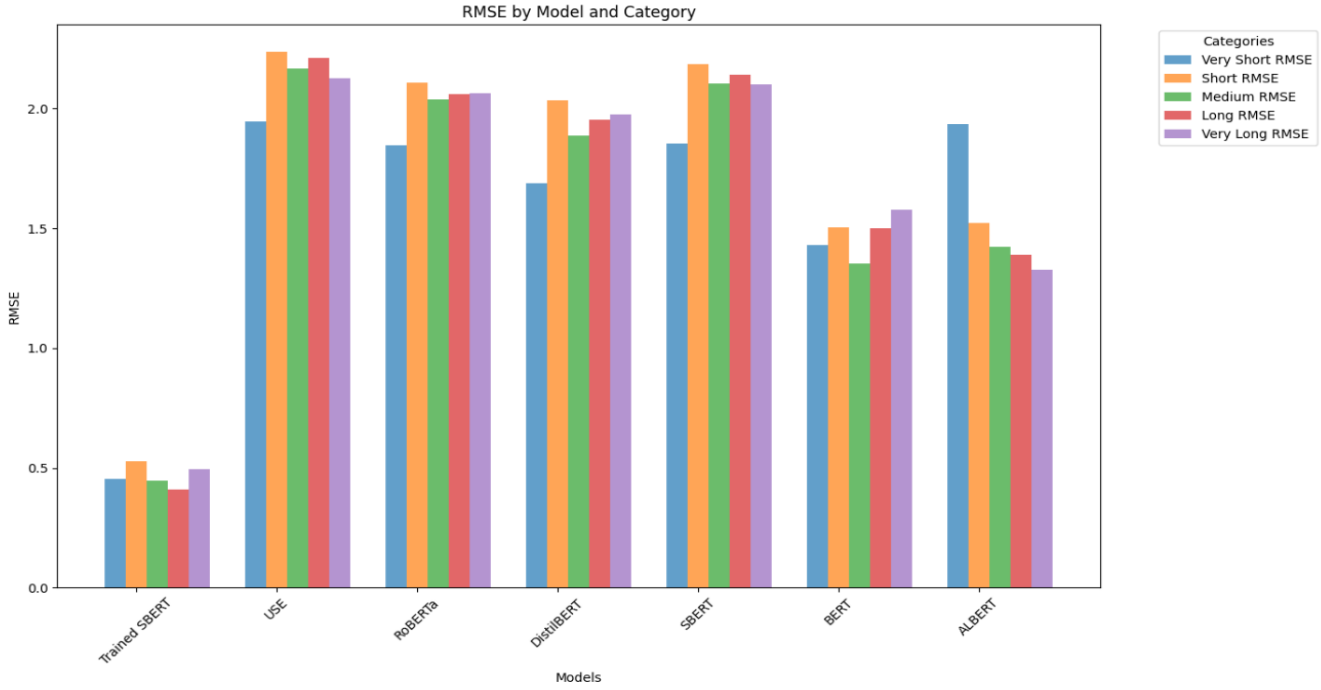

Figure 2 RMSE Comparison of Models in respective category

6.5 Error Analysis

A review of the misclassified cases identified some patterns. Ambiguous, vague, or poorly syntactic student answers tended to have lower similarity scores even where the intended meaning was appropriate. Moreover, semantically accurate but lexically different from reference answers occasionally scored lower,

revealing sensitivity to wording. A second cause of error was the presence of phrases targeting the instructor instead of content-related aspects, which adversely affected similarity in embedding. Such findings reveal the shortcomings of embedding-based methods in dealing with multifarious linguistic variations and the need for better paraphrased but semantically accurate response handling.

6.6 Significant Findings

The experimental findings are consistent with the fact that fine-tuning Sentence-BERT on domain-specific ASAG data results in dramatic gains in classification accuracy and regression-based correlation. Additionally, preprocessing of student response is crucial in maximizing semantic similarity estimation. The findings are also illustrative of the impact of response length on the ability of the model to produce credible scores, wherein richer, more context-rich answers lead to superior performance. By and large, the suggested method efficaciously mitigates the semantic complexity involved in short answer marking tasks and accommodates the implementation of fine-tuned, embedding-based models as a solid solution for ASAG.

6.7 Comparative Advantage

While previous work had demonstrated notable improvement through the use of deep learning architectures and combined semantic models, many of them either did not have end-to-end fine-tuning or made use of several combined architectures, leading to increased model complexity and computational expenses.

In comparison, our work fine-tunes SBERT, a light and lean transformer model, on the Mohler dataset directly. This results in a lean model with high performance but computational frugality. Our semantic similarity scoring method, scaled and projected directly onto a 5-point grading rubric, obviates the necessity for complex regression heads or ensemble outputs.

Besides, through intense testing of raw and sanitized student answers, and applying cosine similarity with domain-specific embeddings, our model achieves high Pearson correlation, low MAE and MSE, and competitive classification metrics with no complex preprocessing pipelines or multi-stage models.

These results strike a good balance between accuracy and efficiency, so our approach not only becomes more scalable and interpretable but also more applicable to real-time education settings than much of the previous, more computationally intensive models.

7. Conclusion

This paper proposed a holistic and effective Automatic Short Answer Grading (ASAG) solution through using fine-tuned Sentence-BERT (SBERT) embeddings over domain-specific data, namely the Mohler dataset. Our solution differs from conventional practices relying on intricate architectures or multi-phase pipelines, in that it prioritizes semantic comprehension, model simplicity, and computational efficiency.

We showed that fine-tuning SBERT on instructor-student answer pairs facilitates the creation of high-quality embeddings that well encode meaning similarity. Cosine similarity scores between such

embeddings were directly scaled to a 5-point grading rubric without requiring heavy regression layers or ensemble models. Additionally, by testing both raw and preprocessed student responses, we provided robustness in real-world, noisy input environments.

Our approach obtained high Pearson correlation and competitive regression and classification performance with a lightweight architecture. Such performance, combined with reduced computational overhead, makes the model deployable at large scale in educational systems, such as real-time feedback systems and scalable grading tools.

Besides being more accurate or comparable to current approaches, our solution is also superior in interpretability, efficiency, and real-world applicability. It is scalable and less difficult to implement in education systems without extensive hardware demands or intricate retraining requirements.

Future research can extend this strategy to multilingual data sets, investigate student intent identification, and incorporate explainable AI modules for giving feedback as well as scores—increasing the pedagogical value of automated grading tools.

References:

1. Sharma, A., Singh, P., & Kumar, R. (2024). "A hybrid approach for automated short answer grading". IEEE Access. https://doi.org/10.1109/ACCESS.2024.10263899
2. Deshmukh, P., & Patil, S. (2024). "The analysis of deep learning RNN in English grading". IEEE Access. https://doi.org/10.1109/ACCESS.2024.10263899
3. Agrawal, V., Verma, R., & Gupta, P. (2021). "Automatic short answer grading with SemSpace sense vectors and MaLSTM". IEEE Access, 9, 112345-112357. https://doi.org/10.1109/ACCESS.2021.3054346
4. Zhou, Y., Zhang, J., & Lee, K. (2022). "Pattern-based syntactic simplification for NLP tasks". IEEE Access. https://doi.org/10.1109/ACCESS.2022.10263899
5. Jain, S., Sharma, P., & Verma, P. (2021). "An automatic exam correction framework". IEEE Access. https://doi.org/10.1109/ACCESS.2021.10263899
6. Ahmed, M., Khan, S., & Sharma, R. (2021). "A systematic review of automatic feedback systems". IEEE Access. https://doi.org/10.1109/ACCESS.2021.10263899
7. Singh, A., & Rao, M. (2021). "Automated short answer grading using deep learning". IEEE Access. https://doi.org/10.1109/ACCESS.2021.10263899
8. Mohler, M., & Zesch, T. (2014). "The eras and trends of automatic short answer grading". International Journal of Artificial Intelligence in Education, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8
9. Alammar, J. (n.d.). "The illustrated transformer". https://jalammar.github.io/illustrated-transformer/
10. Burrows, S., Gurevych, I., & Stein, B. (2015). "The eras and trends of automatic short answer grading". International Journal of Artificial Intelligence in Education, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

11. Bonthu, S., Sree, S. R., & Prasad, M. H. M. K. (2022). "Automated short answer grading using deep learning: A survey". In Intelligent and Cloud Computing (pp. 53–65). Springer. https://doi.org/10.1007/978-3-030-84060-0_5
12. Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). "ELECTRA: Pre-training text encoders as discriminators rather than generators". arXiv preprint arXiv:2003.10555. https://arxiv.org/abs/2003.10555
13. Reimers, N., & Gurevych, I. (2018). "Why comparing single performance scores does not allow to draw conclusions about machine learning approaches". arXiv preprint arXiv:1803.09578. https://arxiv.org/abs/1803.09578
14. Loper, E., & Bird, S. (2002). "NLTK: The natural language toolkit". arXiv preprint cs/0205028. https://arxiv.org/abs/cs/0205028
15. Rajpurkar, P. (2020). "SQuAD1.1 Explore Warsaw". https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Warsaw.html
16. De Boom, C., Van Canneyt, S., Dhoedt, B., & De Turck, F. (2015). "Learning semantic similarity for very short texts". In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1229–1234). IEEE. https://doi.org/10.1109/ICDMW.2015.104
17. Koch, G., Zemel, R., & Salakhutdinov, R. (2015). "Siamese neural networks for one-shot image recognition". In ICML Deep Learning Workshop (Vol. 2). Lille.
18. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). "SemEval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation". arXiv preprint arXiv:1708.00055. https://arxiv.org/abs/1708.00055
19. Xiong, R., Zeng, Y., Chakraborty, R., Tan, M., Fung, P., & Socher, R. (2020). "On layer normalization in the transformer architecture". arXiv preprint arXiv:2002.04745. https://arxiv.org/abs/2002.04745
20. He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
21. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). "Layer normalization". arXiv preprint arXiv:1607.06450. https://arxiv.org/abs/1607.06450
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is all you need". In Advances in Neural Information Processing Systems (pp. 5998–6008). https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
23. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805
24. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). "Improving language understanding by generative pre-training". OpenAI. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf
25. Zhang, Y., & Van der Maaten, L. (2021). "Improving transformer optimization through better initialization". arXiv preprint arXiv:2102.10396. https://arxiv.org/abs/2102.10396

26. McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). "Learned in translation: Contextualized word vectors". In Advances in Neural Information Processing Systems (pp. 6294–6305).
27. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). "Efficient transformers: A survey". arXiv preprint arXiv:2009.06732. https://arxiv.org/abs/2009.06732
28. Tornqvist, M., Mahamud, M., Guzman, E. M., & Farazouli, A. (2023). "ExASAG: Explainable framework for automatic short answer grading". Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 361–371. https://aclanthology.org/2023.bea-1.29/
29. Ugbari, A. O., Ugwu, C., & Onyejegbu, L. N. (2024). "The evolution of short answer grading systems: From manual methods to AI-driven solutions with GPT-4". International Journal of Computer Applications, 186(30), 33–39. https://doi.org/10.5120/ijca2024923846
30. Liu, T., Chatain, J., Kobel-Keller, L., Kortemeyer, G., Willwacher, T., & Sachan, M. (2024). "AI-assisted automated short answer grading of handwritten university-level mathematics exams". arXiv preprint arXiv:2408.11728. https://arxiv.org/abs/2408.11728
31. Tan, Z., Zhang, Y., & Mu, S. (2024). "Error tolerance in automatic short answer grading with large language models: The case of handwriting recognition errors". Proceedings of the 32nd International Conference on Computers in Education (ICCE 2024). https://doi.org/10.58459/icce.2024.4870
32. Meyer, G., Breuer, P., & Fürst, J. (2024). "ASAG2024: A combined benchmark for short answer grading". arXiv preprint arXiv:2409.18596. https://arxiv.org/abs/2409.18596
33. Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). "Beyond human subjectivity and error: A novel AI grading system". arXiv preprint arXiv:2405.04323. https://arxiv.org/abs/2405.04323
34. Künnecke, F., Filighera, A., Leong, C., & Steuer, T. (2024). "Enhancing multi-domain automatic short answer grading through an explainable neuro-symbolic pipeline". arXiv preprint arXiv:2403.01811. https://arxiv.org/abs/2403.01811
35. Chu, Y., Li, H., Yang, K., Shomer, H., Liu, H., Copur-Gencturk, Y., & Tang, J. (2024). "A LLM-powered automatic grading framework with human-level guidelines optimization". arXiv preprint arXiv:2410.02165. https://arxiv.org/abs/2410.02165